Review of Educational Research
Fall 1996, Vol. 66, No. 3, pp. 323–340

Effects of Multigrade and Multi-Age Classes Reconsidered

Simon Veenman
University of Nijmegen

In response to "Cognitive and Noncognitive Effects of Multigrade and Multi-Age Classes: A Best-Evidence Synthesis" (Veenman, 1995), Mason and Burns (1996) report that their research and review of the literature has led them to conclude that multigrade classes have a slightly negative effect on student achievement. They argue, moreover, that multigrade classes generally have better students and perhaps better teachers and that this selection bias masks the negative effects of less effective instruction in multigrade classes. In this rejoinder, a reanalysis, based on meta-analytic procedures, of the available multigrade and multi-age studies shows the average weighted effect sizes to be essentially zero or close to zero. For all analyses, the confidence intervals around the average effect sizes included zero. These results provide little support for the assumption that the quality of instruction in multigrade classes is lower than in single-grade classes. Between-study differences revealed that favorable conditions for classroom instruction, the country of publication, the locality and socioeconomic status of the school, the grade level of the students, and the number of years spent in multigrade classes need the attention of investigators in future research into the effects of multigrade classes.

I appreciate the thoughtful reactions of Mason and Burns (1996) to my review article "Cognitive and Noncognitive Effects of Multigrade and Multi-Age Classes: A Best-Evidence Synthesis" (Veenman, 1995). My critics argue that I have an implicit bias toward multi-age and cross-grade grouping and that this bias colors my interpretation of the findings and leads me to a conclusion that favors multigrade classes. First, I should point out that I did not begin my research with any particular preference for multigrade or multi-age classes. In 1981, I was (in my spare time) the chair of an elementary school board. Due to declining enrollments and uneven class sizes, the teachers and the school board decided to change the single-grade school organization to a multigrade school organization. During three parent evenings, I had to defend the decision of the teachers and the school board. I was challenged by the parents again and again to prove that the multigrade classroom organization did not harm their children's learning, and I had no clear answer at that time. These evenings marked the starting point of my research into multigrade classes.

My colleagues at the University of Nijmegen and I began collecting observational data in multigrade classes in schools that had been forced by declining enrollments to combine children from two (or more) grade levels into a single classroom, and compared these data to data from single-grade classes. We also administered achievement tests for reading, language, and mathematics. Finally, we interviewed the teachers in the multigrade classes in order to identify any problems and concerns. Our data did not show the students in the multigrade classes to achieve significantly less than the students in the single-age classes. Also, the students in the multigrade classes and the students in the single-grade classes did not spend their learning time differently. The interview data showed the teachers in the multigrade classes to be less satisfied with their jobs than their counterparts in single-age classes as a result of the heavy teaching load and demands for classroom management. We concluded that the difficulties teachers face in multigrade classes are centered around five problem areas: (a) the efficient use of instructional time, (b) the design of effective instruction, (c) classroom management, (d) the organization of independent practice or learning, and (e) the formulation of clear and collectively agreed-upon goals for making the multigrade school work (Kral, 1995; Lem, Veenman, & Voeten, 1990; Veenman, Lem, & Voeten, 1988; Veenman, Lem, & Winkelmolen, 1985; Veenman, Voeten, & Lem, 1987).

The next step was to design a staff development program for teachers in multigrade classes. In three studies, we assessed the short-term and long-term effects of this program. Based on pretraining and posttraining classroom observation, these studies revealed a significant treatment effect for the time-on-task levels of the students in the multigrade classes and for the instructional and management skills of the teachers. However, no significant differences in achievement were found between the students in classes with trained multigrade teachers and the students in classes with untrained single-grade teachers (Roelofs, 1993; Roelofs, Raemaekers, & Veenman, 1991; Roelofs, Veenman, & Raemaekers, 1994; Veenman & Raemaekers, 1995).

As these studies were being conducted, Dutch school board members, school principals, teachers, and parents regularly raised the same question as in 1981: What does the research say about the effects of multigrade classes on student learning? After our training studies had been completed, I tried to answer this question by examining the available research. I had no particular bias toward the multigrade or multi-age grouping but was simply curious about the cognitive and noncognitive effects of multigrade versus single-grade classes across a variety of English-speaking and non-English-speaking countries. In examining the literature, I found that various reviewers (e.g., Miller, 1990, 1991; Pratt, 1986) were often lumping multigrade and multi-age classes together when, in fact, these two forms of classroom organization were grounded in different motives. I was therefore careful to distinguish between these two forms of classroom organization in examining their effects. Before responding to my critics in greater detail, I would like to present the results of a reanalysis, using meta-analytic procedures, of the data reported in my best-evidence synthesis.

A Meta-Analysis of Multigrade and Multi-Age Studies

The 56 studies on which my best-evidence synthesis (Veenman, 1995) was based had to meet a set of a priori criteria with respect to germaneness and methodological adequacy (cf. Slavin, 1986). All of the studies included in the best-evidence synthesis had to involve a comparison of the cognitive or noncognitive effects of multigrade classes and single-grade classes, or multi-age classes and single-age classes, in elementary schools. The methodological requirements for inclusion were as follows:

(1) Experimental and control groups were compared.
(2) Standard measures were used for cognitive and noncognitive outcomes.
(3) Samples showed initial comparability.
(4) Multigrade or multi-age grouping had been in place for at least 1 year.
(5) Multigrade or multi-age grouping was in regular (i.e., nonspecial) elementary classrooms.
(6) Teachers in the experimental groups had not been trained on the dependent measures.
(7) At least two experimental and two control teachers were involved in each study.

In meta-analyses, it is common to include all conceptually relevant studies and quantitatively examine the possibility of specific methodological features being related to study outcomes. If the study results are found to be related to differences in methodological quality, then the conclusions based on the methodologically sound studies are the ones to be believed (cf. Cooper & Dorr, 1995). Following this practice, four studies and one substudy excluded from the best-evidence synthesis were included in the meta-analysis. These were studies by Eames (1989) and MacDonald and Wurster (1974) that each involved only one experimental and one control teacher, training studies by Roelofs (1993) and Veenman and Raemaekers (1995), and the reading achievement part of a study by Hoen (1972) that involved incomparable samples. The data for the meta-analysis were further supplemented with three recent studies not mentioned in the best-evidence synthesis: Pawluk (1992), Knuver (1993), and Doolaard (1996). The study by Pawluk arrived too late to be included in the best-evidence synthesis. In this study, the academic achievement of students enrolled in Grades 5–8 in Seventh-Day Adventist parochial schools in Oregon and Washington was examined. The studies by Knuver and Doolaard included a national random sample of Dutch elementary schools, but the two sets of data had yet to be analyzed with respect to the cognitive and noncognitive effects of multigrade classes versus single-age classes. The study by Knuver was focused on noncognitive effects for students in Grade 6; the study by Doolaard was directed at cognitive and noncognitive effects for students in Grade 4. The main characteristics of all of these supplementary studies, not included in the best-evidence synthesis but included in the present meta-analysis, are presented in Table 1. The meta-analysis thus included 51 multigrade studies (45 studies concerning the cognitive effects and 19 studies concerning the noncognitive effects of multigrade classes) and 12 multi-age studies (11 studies concerning the cognitive effects and 11 studies concerning the noncognitive effects of multi-age classes). It should be noted that studies containing reports of both cognitive and noncognitive effects were counted twice.

Computation and Analysis of Effect Sizes

The procedures used in the meta-analysis followed those of Hedges and Olkin (1985). The effect sizes, g, in the best-evidence synthesis were generally computed as the difference between the experimental and control means divided by the control standard deviation (Glass, McGaw, & Smith, 1981). When means or standard deviations were missing, the effect sizes were estimated from ts, Fs, or

TABLE 1
Main characteristics of supplementary studies

Article	Grades	Location	Sample size		Study type	*ES*
			MG	SG		
		Multigrade classes: Cognitive effects				
Doolaard, 1996	4	Netherlands	1,373	2,094	Study using random sample	–.05
Eames, 1989	4	Connecticut, USA	22	22	Matched study lacking evidence of initial equality (without adjustment for pretest differences)	.32
Hoen, 1972	5	Vancouver, Canada	12	46	Matched study lacking evidence of initial equality (with adjustment for pretest differences)	.00
MacDonald & Wurster, 1974	1	Arizona, USA	20	20	Matched study lacking evidence of initial equality (without adjustment for pretest differences)	–.18
Pawluk, 1992	5-8	Oregon, Washington, USA	172	116	Matched study lacking evidence of initial equality (with adjustment for pretest differences)	–.25
Roelofs, 1993	1–6	Southern Netherlands	1,145	342	Matched study lacking evidence of initial equality (with adjustment for pretest differences)	–.11
Veenman & Raemaekers, 1996	1–6	Southern Netherlands	632	338	Matched study lacking evidence of initial equality (without adjustment for pretest differences)	.07
		Multigrade classes: Noncognitive effects				
Doolaard, 1996	4	Netherlands	1,373	2,094	Study using random sample	–.03
Knuver, 1993	6	Netherlands	998	2,883	Study using random sample	–.001

Note. MG = multigrade classes, SG = single-grade classes, *ES* = effect size (*g*-index).

*p*s, or other statistics (see Glass et al., 1981). When the results for a particular outcome measure were either not reported or reported as nonsignificant, the effect size was set to 0.00 for the meta-analysis. For each analysis, the results were calculated first with the 0.00 values included and then with the 0.00 values omitted.

When more than one *g*-index could be calculated for a single dependent measure (e.g., reading, language, mathematics, or social studies), these were averaged within the samples in order to ensure independent effect sizes (cf. Cooper & Dorr, 1995). When a single study reported separate results for cognitive and noncognitive effects, two overall effect sizes were calculated, one for the dependent variable *cognitive effect* and one for the dependent variable *noncognitive effect.* For this reason, the total number of effect sizes exceeds the number of studies. In order to satisfy the independence assumption of meta-analytic statistics (Hedges & Olkin, 1985), only one effect size per study was entered into each analysis.

After all of the effect sizes had been calculated, the analyses were conducted using the computer program Meta-Analysis Programs (Schwarzer, 1991). A weighted average effect size (or d-index) across a series of studies was then calculated by multiplying each g-index by the inverse of its variance and dividing the sum of these products by the sum of the weights (Hedges & Olkin, 1985). Functionally, this procedure gives proportionally greater weight to effect sizes based on larger samples. Weighted average effect sizes are more precise estimates of population values than unweighted ones. A 95% confidence interval was then calculated for this weighted estimate.

To determine whether each set of effect sizes in a sample shared a common effect size (i.e., was consistent across studies), a homogeneity statistic, Q, was calculated. Q has an approximate chi-square distribution with $k - 1$ degrees of freedom, where k is the number of effect sizes (Hedges & Olkin, 1985). This procedure was used to test whether sampling error alone accounted for variation in effect sizes or whether specific features of the studies, samples, or outcome measures also contributed to this variation. When the samples were not homogeneous, the studies could be classified according to potentially important characteristics, such that the effect sizes within categories were homogeneous. This strategy was used to examine the effects of a number of different subclasses of studies. For purposes of the present analyses, samples were considered homogeneous at $p < .01$. In all of the analyses, the random effects model was used; that is, the study sample was presumed to be a sample from a hypothetical collection (or population) of studies (Hedges, 1994).

As a supplementary analysis, homogeneity was attained by removing outliers. That is, studies were omitted when they provided estimates that were inconsistent with those from other studies. Outliers in each set were identified by performing a (disjoint) cluster analysis (Hedges & Olkin, 1985; Schwarzer, 1991) and by inspecting box plots generated with SPSS for Windows. In the overall analyses, outliers were both included and excluded. In subsequent analyses, those studies containing outliers were omitted.

Identification of Outliers

In the studies that reported effect sizes for the cognitive effects of multigrade classes, five outliers were identified. Three of these five were found in studies conducted in developing countries. The first two were positive outliers found in the studies conducted by Jarousse and Mingat (1991, 1992) in Africa (g-index = +.50 and +.42, respectively), and the third was a negative outlier found in a study conducted by Rowley (1992) in Pakistan (g-index = –.36). These effects could be due to differences in the educational systems in developing versus developed countries. According to Fuller (1987), the school institution in the third world often operates within communities where the commitment to written literacy and numeracy is relatively recent. This means that a school of even modest quality may significantly influence academic achievement. School factors such as expenditure per student, availability and use of textbooks, availability of a school library, teacher quality, length of instructional program, child-rearing beliefs, and child's nutritional status may also play a greater role in developing countries than in industrialized countries (Fuller, 1987). The class sizes in the studies by Jarousse and Mingat (1991, 1992) fell outside the normal ranges; in Togo and Burkina Faso, they varied from 23 to 150 students. In the study by Rowley (1992) in Pakistan, most of the highly trained teachers were found in the single-grade classes. In addition, when teachers were assigned to teach in an area that was not their own ethnic or linguistic area, they had to make use of student translators. The factors associated with schools in developing countries versus developed countries may thus moderate the effects of multigrade classes; these factors certainly demand attention in future research.

The remaining two outliers were found in the studies by Fippinger (1967; *g*-index = –.44) and Martens (1954; *g*-index = –.61). These studies differed from the other studies in that the grouping effects were confounded with location effects: the single-grade schools in urban areas were found to be populated by students with higher intellectual abilities than the multigrade schools in rural areas. In these two studies, neither socioeconomic status nor intelligence was controlled for. The differences between the multigrade classes in the rural and (sub)urban areas will be examined in a subsequent analysis.

Six outliers were identified in those studies that reported effect sizes for the noncognitive effects of multigrade classes. For this data set, two clusters emerged in addition to the main cluster. The studies with positive outliers (*g*-indexes ranging from +.44 to +.28) were conducted by Harvey (1974), Chace (1961), Carter (1973), Rehwoldt and Hamilton (1957), and Knörzer (1985), and they could be grouped together because their school settings or classroom conditions might have favored multigrade classes. The study by Harvey was conducted in a rural county in Virginia where people had had less-than-average schooling for rural Virginians and were at a disadvantage with regard to income when compared to income levels for other rural localities. The positive outlier was produced by the high self-concept scores for the lower-class children in the three multigrade classes with kindergartners and first graders. The positive effect can thus be explained by the greater influence of school in general within disadvantaged areas relative to advantaged areas (cf. Fuller, 1977). The school containing multigrade classes in the study by Chace was a campus laboratory school for a teacher training college located in a rural environment. The teachers were encouraged by the administration to disregard grade limitations in favor of individual development. One multigrade class contained students who had been placed in it at the request of their parents. The schools included in the study by Carter had composite scores on standardized achievement tests above the national grade-level norms. The study by Rehwoldt and Hamilton was confined to one school and to children whose parents had requested their placement in multigrade classrooms. Finally, the study by Knörzer was restricted to small schools in rural areas, and the multigrade classes were smaller than the comparable single-grade classes.

The study conducted by Zabolotney (1983) with a negative outlier (*g*-index =

–.39) constitutes a cluster of its own. This study differed from others in that it was conducted with students from low- to medium-income families in rural multigrade Seventh-Day Adventist schools in Arkansas. Half of the multigrade classes consisted of Grades 1 through 4, and half had combinations ranging from Grades 3 and 4 to Grades 1 through 6. All of the classes were taught by one teacher.

In sum, the six studies identified as outliers in the data set for the noncognitive effects of multigrade classes differed from the other studies in that particular groups of students and/or particular schools were used. For instance, one study involved rural students from families with less-than-average schooling. In others, only high-achieving students were considered, or parents volunteered their children for multigrade classes. On the school level, these studies involved Seventh-Day Adventist schools, small rural schools with small class sizes or only one teacher, and a campus laboratory school. The studies also used very different instruments to assess noncognitive outcomes. The characteristics of the students and schools, the role of the parents, and the differences in the noncognitive tests should therefore serve as moderator variables in future analyses of the noncognitive effects of multigrade versus single-grade classes.

No outliers were identified in the data set for the cognitive effects of multi-age classes. One positive outlier (*g*-index = +.58) in the data set for the noncognitive effects of multi-age classes was found in a study conducted by Givens (1972). This study differed from others in that the multi-age grouping was practiced in a demonstration school associated with a local university. This school also featured team teaching, an open space concept, and individualized instruction. Admission to the school was by application only.

Meta-Analysis of Effect Sizes

The results of the overall meta-analysis of the effect sizes are presented in Table 2. Because average scores for single-grade classes were subtracted from average scores for multigrade classes, positive values in this table indicate that the multigrade classes scored higher on the average than the single-grade classes. The analyses were conducted for the following four data sets: (a) studies reporting cognitive effects of multigrade classes, (b) studies reporting noncognitive effects of multigrade classes, (c) studies reporting cognitive effects of multi-age classes, and (d) studies reporting noncognitive effects of multi-age classes. First, an analysis was conducted for all of the studies providing information with regard to the effects of multigrade/multi-age classes (see the rows labeled *All studies* in Table 2). Second, an analysis was conducted for all of the studies for which effect sizes could be estimated (see *All known effects* in Table 2). Third, when the samples were not homogeneous, the outliers were removed (with zeros initially included and later excluded). And finally, an analysis was conducted for the studies included in my best-evidence synthesis, reported last year in the *Review of Educational Research* (Veenman, 1995; see *Best-evidence studies* in Table 2).

An examination of the weighted average effect sizes in Table 2 indicates that including and excluding outliers and zeros do not drastically alter the mean effect sizes. Overall, the results show that there is no significant difference in either cognitive or noncognitive learning outcomes between multigrade/multi-age grouping and single-grade/single-age grouping. For all of the analyses, the confidence intervals around the average effect sizes included 0.00.

TABLE 2

Mean d*-indexes for overall cognitive and noncognitive effects of multigrade versus single-grade classes and multi-age versus single-age classes*

Analysis	*k*	Sample size	*d*-index	Confidence interval	*Q*
		Multigrade classes: Cognitive effects			
All studies (zeros and outliers included)	45	73,225	+.001	–.05 to .05	578.80*
All known effects (zeros excluded, outliers included)	40	69,913	–.001	–.06 to .06	578.68*
Outliers excluded, zeros included	40	62,467	–.002	–.02 to .01	266.87*
Outliers excluded, zeros excluded	35	59,155	+.01	–.03 to .04	266.87*
Best-evidence studies (zeros excluded, outliers included)	34	63,621	+.01	–.07 to .08	564.14*
		Multigrade classes: Noncognitive effects			
All studies (zeros and outliers included)	19	16,309	+.07	–.02 to .16	61.50*
All known effects (zeros excluded, outliers included)	15	15,071	+.08	–.03 to .20	61.28*
Outliers excluded, zeros included	13	14,556	–.003	–.04 to .03	17.25
Outliers excluded, zeros excluded	9	13,318	–.004	–.04 to .03	17.25
Best-evidence studies (zeros excluded, outliers included)	13	7,723	+.11	–.02 to .24	54.83*
		Multi-age classes: Cognitive effects			
All studies (zeros included, no outliers)	11	4,142	–.05	–.15 to .04	12.35
All known effects/best-evidence studies (zeros excluded, no outliers)	8	2,390	–.08	–.21 to .06	11.01

TABLE 2 *(continued)*

Analysis	*k*	Sample size	*d*-index	Confidence interval	*Q*
		Multi-age classes: Noncognitive effects			
All studies (zeros and outlier included)	11	4,104	+.08	−.03 to .20	17.58
All known effects/best-evidence studies (zeros excluded, outlier included)	8	2,352	+.13	−.02 to .27	16.71
Outlier excluded, zeros included	10	4,004	+.03	−.04 to .11	10.62
Outlier excluded, zeros excluded	7	2,252	+.06	−.04 to .16	10.25

Note. Best-evidence studies are those included in the best-evidence synthesis published last year in *Review of Educational Research* (Veenman, 1995).
$^*p < .01$.

Cognitive Effects of Multigrade Classes: Further Between-Study Differences

The results of the homogeneity analyses show homogeneity to be attained for the subsets of studies concerned with the noncognitive effects of multigrade classes and the cognitive and noncognitive effects of multi-age classes after removal of the observed outliers. No further analyses were performed on these homogeneous subsamples. A homogeneity analysis for the subset of studies concerned with the cognitive effects of multigrade classes (zeros included) revealed that there was considerably more variability in the 40 individual *d*-indexes after the exclusion of the outliers than would be predicted by sampling error alone, $Q(39) = 266.87, p < .001$. In order to examine this variation further, these studies were grouped according to the following distinctions: country of publication, type of study, source of publication, location of the school, socioeconomic status of the school, grades under study, and number of years students spent in multigrade classrooms. With the exception of the country and the source of publication, all of these distinctions were also considered in the best-evidence synthesis. The results of the between-study analyses are presented in Table 3.

With regard to the countries in which the studies were conducted, three subclasses were formed: studies from the United States, studies from Canada, and studies from Europe. As noted before, studies from developing countries were excluded in this breakdown, because three of the four studies conducted in the third world were identified as outliers. The results of the between-study analysis showed the country of publication to predict a significant amount of variance in effects, $Q(2, N = 39) = 17.2, p < .001$. The results in Table 3 show a small positive effect for studies conducted in the United States (*d*-index = .05) and Canada (*d*-index = .08) and a small negative effect for studies conducted in Europe (*d*-index = −.05). This difference between the two continents may reflect international differences in school factors (e.g., school organization, student-teacher ratios,

TABLE 3
Results of between-study analyses for the cognitive effects of multigrade versus single-grade classes

Subclasses	*k*	Sample size	*d*-index	Confidence interval
Country of publication[a]				
United States	22	20,144	.05	.02 to .08
Canada	3	4,883	.08	.03 to .14
Europe	14	34,407	–.05	–.11 to .01
Type of study[b]				
Matched study with evidence of initial equality	9	10,998	.05	.01 to .09
Study using random samples	13	41,126	–.01	–.08 to .07
Matched study lacking evidence of initial equality	18	10,343	–.02	–.06 to .02
Source of publication[c]				
Book	2	1,523	.000	–.11 to .11
Doctoral dissertation	16	23,740	.001	–.06 to .06
Journal article	11	8,568	.08	.04 to .13
Research report	11	28,636	–.05	–.08 to –.03
Locality[d]				
(Sub)urban	17	19,050	.06	.03 to .09
Rural	9	6,034	.10	.05 to .15
Socioeconomic status[e]				
Upper/middle class	17	18,681	.06	.03 to .09
Lower class	5	3,844	.09	–.01 to .20
Grade level[f]				
Lower grades (K–2)	9	8,038	.06	.02 to .11
Intermediate grades (3–4)	20	37,649	.01	–.02 to .05
Higher grades (5–6)	5	7,799	–.08	–.20 to .05
Number of years spent in multigrade classes[g]				
1 year	20	37,612	–.02	–.05 to –.002
2 years	6	1,891	.01	–.08 to .10
4 years	4	1,778	–.02	–.12 to .07
6 years	8	17,729	–.02	–.08 to .05

Note. Positive values indicate higher means for multigrade samples than for single-grade samples; *k* = number of studies contributing to estimate. Outliers excluded.
[a]$Q(2) = 17.24, p < .01$; [b]$Q(2) = 1.58, p > .40$; [c]$Q(3) = 1.53$, p > .60; [d]$Q(1) = 4.88, p > .02$; [e]$Q(1) = 3.28, p > .05$; [f]$Q(2) = 18.89, p < .01$; [g]$Q(3) = 18.47, p < .01$.

class size, grouping practices, classroom resources, parental involvement), teacher characteristics (e.g., training, experience), instructional strategies (e.g., activities and methods), aspects of the curriculum (e.g., content covered), and measuring instruments (e.g., forms and aspects covered; cf. Lundberg & Linnakylä, 1993). The country of publication thus deserves attention in future research and syntheses.

As an index of the methodological quality of the studies, they were grouped into three classes: matched studies with evidence of initial equality, studies using random samples, and matched studies lacking evidence of initial equality. The first two groups of studies were considered higher in methodological quality than the last group of studies. The results of this between-study analysis showed that study quality did not predict a significant amount of variance in effects, $Q(2, N = 40) = 1.6, p = .45$.

In order to examine the possibility of a publication bias, the studies were grouped into four classes: books, doctoral dissertations, journal articles, and research reports. As can be seen in Table 3, the source of publication did not significantly influence the magnitude of the effect sizes, $Q(3, N = 40) = 1.5, p = .68$.

The studies were next divided into those that examined the cognitive outcomes of multigrade classes in (sub)urban areas and those that examined these outcomes in rural areas. The locality of the school did not predict a significant amount of variance in effects, $Q(1, N = 26) = 4.9, p = .03$.

In order to examine the relation between multigrade versus single-grade grouping and the socioeconomic status of the schools, the studies were divided into those that were described by the researchers as containing both middle- and upper-class students and those containing only lower-class students. As displayed in Table 3, the results of this between-study analysis showed that the socioeconomic status of the schools did not predict a significant amount of variance in effects, $Q(1, N= 22) = 3.3, p = .07$.

A significant amount of variance in effects was found for grade level, $Q(2, N = 34) = 18.9, p < .01$. The average weighted *d*-indexes across these studies showed a diminishing positive effect of multigrade grouping. The studies concerned with lower grades (K–2) show a small positive effect for multigrade grouping; the effect for intermediate grades (3–4) is essentially zero; and the studies directed at higher grades (5–6) show a small negative effect. It can be hypothesized that the intellectual abilities of the students grow further apart as the students enter the higher elementary grades, and that the teachers in single-age classes can deal with these differences better than teachers in multigrade classes. It should be noted that a single study could be responsible for more than one effect size when data for more than one grade level were available.

The final between-study analysis revealed significant differences in the average effect sizes for the number of years students spent in multigrade classes, $Q(3, N = 38) = 18.5, p <.01$. As can be seen in Table 3, however, all *d*-indexes for the four groups of studies were essentially zero, so the observed differences are negligible. It should be noted that those studies lacking information with regard to the number of years students spent in multigrade classes were placed in the minimal category of *1 year.*

The final moderator examined in this meta-analysis concerned the year in which a study was published. This variable was not, however, found to be significantly related to the magnitude of the effect sizes, $r(40) = -.05$, n.s.

Although some of the between-study analyses show significant differences for potentially relevant study characteristics, these differences were generally found to be very small. The weighted average *d*-indexes were essentially zero or near zero. In most instances, the confidence interval included 0.00. The country of publication, the locality and the socioeconomic status of the school, the grade level of the students, and the number of years students spent in multigrade classes should nevertheless be examined as potential mediators of differences between multigrade and single-grade grouping in future research.

In Reply

The claim by Mason and Burns (1996) that I fell victim to the advocacy of multi-age classes is based on a section of my review devoted to the supposed advantages of multi-age grouping. Mason and Burns suggest that I should have devoted an equivalent section to the disadvantages of multi-age classes and concomitant advantages of multigrade classes. My critics also argue that I paid little attention to important teacher and principal concerns with regard to multigrade classes and therefore undervalued the role of such factors.

In reviewing the research studies on multigrade and multi-age classes, I found that the studies on multi-age teaching generally considered the value of this form of classroom organization and suggested that it was superior to single-age grouping. For this reason, I devoted a section of my article to the presumed advantages of multi-age classes, and to the question of whether such claims can be substantiated by empirical research. In most of the studies of multigrade classes, little or no attention was devoted to the supposed advantages of this specific form of classroom organization. This is because the reason for combining grades is usually declining enrollments or uneven class sizes. Superior effects are rarely claimed, and when such cognitive or noncognitive advantages are claimed, they are usually based on the literature with regard to multi-age grouping. In this way, I suggest, schools make a virtue out of the necessity of multigrading.

In the section of my review devoted to the problems and concerns associated with multigrade grouping, I clearly emphasized the importance of the teachers', principals', and parents' perceptions of multigrade classes. I referred to a number of surveys conducted in the United States, Canada, Australia, the Netherlands, Switzerland, and the United Kingdom and concluded that these studies conducted at different times and in different countries revealed some common problems and concerns with regard to multigrade classes: "lack of time for teaching the required content, a greater workload, lack of time for individual attention and remediation, lack of adequate classroom management skills, lack of adequate preparation during teacher training, inadequate materials, and parental concerns about the academic achievement of their children" (Veenman, 1995, p. 324). Mason and Burns suggest that these concerns and problems may have a negative effect on student outcomes in multigrade classes. However, a reanalysis of the multigrade studies shows an overall weighted average effect size of essentially zero or near zero for both cognitive and noncognitive effects (see Table 2).

Although Mason and Burns essentially accept the evidence as showing "no significant differences" in achievement between multigrade and single-grade classes, and between multi-age and single-age classes, they claim that their research and review of the literature show multigrade classes to have a slightly negative effect on achievement. Unfortunately, this claim is not substantiated with data from the studies they reviewed. I suspect that their conclusions are mainly based on studies conducted in the United States and Canada. As the results in

Table 3 show, the weighted average *d*-index for the 22 studies conducted in the United States equaled +.05, and the weighted average *d*-index for the 3 Canadian studies equaled +.08 (unweighted average *g*-indexes were +.01 and +.07, respectively). For the 14 European studies, however, the weighted average *d*-index equaled –.05 (unweighted average *g*-index = –.04). In other words, a very small negative effect has been found only for the studies conducted in Europe and not for the studies conducted in the United States and Canada.

Based on their conclusion that multigrade classes have a slightly negative effect on achievement, Mason and Burns argue that multigrade classes nevertheless have generally better students and perhaps better teachers and that this selection bias masks the negative effects of less effective instruction in multigrade classes. This interpretation is based on several interview studies with teachers in multigrade classes located in California's year-round schools. In examining the studies from 12 countries, I also paid close attention to just how the multigrade classes were formed and actually looked for clues for a selection bias. In 4 of the 51 studies concerned with multigrade classes, a possible selection bias was explicitly mentioned (Adair, 1978; Brown & Martin, 1989; Knight, 1938; Spratt, 1986). On the basis of the characteristics of the studies in my review, however, I find it premature to say that multigrade classes generally have better students and perhaps better teachers.

In discussing the finding of no difference in achievement between multigrade and single-grade classes, I argue that such selection criteria will most likely be applied in (sub)urban schools with brighter, more independent, and perhaps more motivated students placed in the multigrade classes. In most of the studies conducted in rural areas where student selection is simply not an option (e.g., Nieminen's, 1979, and Jokinen's, 1979, studies in Finland), no significant differences in student achievement have been found between multigrade and single-grade classes, which suggests that the form of grouping itself does not significantly affect student achievement either positively or negatively. As Table 3 displays, the differences between multigrade classes in (sub)urban and rural schools were very small: the weighted average *d*-indexes were +.06 for schools located in (sub)urban areas (unweighted average *g*-index = +.04) and +.10 for schools located in rural areas (unweighted average *g*-index = –.03).

As noted before, I agree that teaching in multigrade classes is more difficult than teaching in single-grade classes. I cannot, however, endorse the (unsubstantiated) conclusion of my critics that a (still yet to be proven) selection bias masks the negative effects of putatively inferior instruction in multigrade classes. In reviewing the studies concerned with multigrade classes, I searched for descriptions of the instructional practices in this particular classroom structure. As noted in the best-evidence synthesis, most of the studies provided no information about the instructional practices in the multigrade classes. When it was available, I reported this information, which showed that two-group teaching was generally used for the basic subjects. This finding corresponds to the recently reported results of an interview study by Mason and Burns (1995), in which the teachers in recently formed multitrack year-round schools in California reported using a mixed approach for teaching multigrade classes: formation of two groups for the basic subjects and use of a whole-class format for the other subjects. For the two-group subjects, the teachers typically alternated between recitation and seatwork for the two groups. The use of such a mixed approach does not show the quality of instruction in multigrade classes to be lower than in single-grade classes, however, and the results in Table 2 provide little ground for such an assumption.

Referring to the studies by Slavin (1987) and Gutiérrez and Slavin (1992) on effective grouping of students, I pointed to the positive effects of cross-grade grouping. I also pointed to the positive effects of cooperative learning and reciprocal teaching. I did not assert that these organizational and instructional approaches would automatically produce more effective multigrade classes. Rather, I concur with Mason and Burns's observation that these approaches have yet to be tested experimentally in a multigrade context. I assume that my critics join me in encouraging the comparison of the cognitive and noncognitive effects of alternative grouping arrangements and instructional practices in multigrade settings.

In summarizing the results of the cognitive effects of multigrade grouping in my best-evidence synthesis, I restricted myself to the basic skills of reading, mathematics, and language because most of the studies I reviewed were limited to these skills. Mason and Burns assume that multigrade teachers pressed for instructional time and the mastery of basic skills will neglect science, social studies, and other subjects. I did not report science and social studies achievement separately because the number of studies that could contribute to reliable estimates of the effect sizes for these subjects was small. The effect sizes for the small number of available studies were nevertheless included in the tables in my best-evidence synthesis. These outcomes were used by Mason and Burns to confirm their assumption that teachers in multigrade classes neglect subjects such as science and social studies, and a median effect size of –.16 is reported for these subjects. In my reanalysis of the data using meta-analytic procedures, I found the following weighted average *d*-indexes: (a) for science –.19 (6 studies, N = 5,989), (b) for social studies –.25 (4 studies, N = 3,970), and (c) for English as a foreign language .04 (3 studies, N = 9,429). The tests for the homogeneity/heterogeneity of effect sizes for these subsamples revealed that the studies were significantly heterogeneous ($p < .01$). The largest effect size (*g*-index = –.91) was found in an atypical study by Dodendorf (1983), who compared students in a two-room rural schoolhouse with single-grade urban students. The results of this reanalysis for science and social studies warrant further investigation in order to test Mason and Burns's assumption that teachers in multigrade classes will neglect other subjects when pressed for time.

On a different front, my critics hold that the categorization of the studies by Junell (1971) and Rehwoldt and Hamilton (1957) as multigrade studies violates my distinction between multigrade and multi-age classes. I classified these two studies as multigrade studies because the investigators consistently define their classes as "multigrade classes." In the study by Junell, the junior high school under study "received all children from a true multigraded elementary school" as well as "all children from two regularly-graded schools" (p. 7). No further information was provided on the characteristics of these elementary schools. In the study by Rehwoldt and Hamilton, it was argued that the "administrative task of placing and equalizing teacher-pupil load will be greatly simplified" (p. 24) by multigrading, which suggests an administrative basis for the adoption of this particular form of classroom organization. I nevertheless agree that the classification of the study by Rehwoldt and Hamilton as multigrade may be open to debate.

In closing, the purpose of my review was not to encourage policymakers and practitioners to adopt the multigraded form of classroom organization more frequently. Policymakers and practitioners should always proceed with caution in the application of research findings, and should not base policy decisions on research findings alone. School board members, school principals, and teachers should take into account not only the findings of research but also the significance of these findings for their own schools (e.g., the distribution of students across grade levels, class size per teacher, work load, teacher commitment and experience, and the concerns and wishes of the parents). I think I sufficiently depicted the concerns of school principals, teachers, and parents with regard to multigrade classes in my review. Rather than restricting myself to the conclusion that there is no difference in achievement between multigrade and single-grade classes, I also offered a number of explanations for this finding and identified issues for further research. Mason and Burns quite correctly note that field experiments that compare achievement and affect in multigrade and single-grade classes are critical—a point which I also argued in the conclusion to my best-evidence synthesis.

References

Adair, J. H. (1978). *An attitude and achievement comparison between kindergarten and first grade children in multi and single grade classes.* Unpublished doctoral dissertation, Boston College.

Brown, K. G., & Martin, A. B. (1989). Student achievement in multigrade and single grade classes. *Education Canada, 29*(2), 10–13, 47.

Carter, J. B. (1973). *A study of the effects of multi-grade grouping on the attitudes toward self, peers, and school of selected third and fifth grade students.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Chace, E. S. (1961). *An analysis of some aspects of multiple-grade grouping in an elementary school.* Unpublished doctoral dissertation, University of Tennessee.

Cooper, H., & Dorr, N. (1995). Race comparisons on need for achievement: A meta-analytic alternative to Graham's narrative review. *Review of Educational Research, 65,* 483–508.

Dodendorf, D. M. (1983). A unique rural school environment. *Psychology in the Schools, 20*(1), 99–104.

Doolaard, S. (1996). *Leren in combinatiegroepen* [Learning in combination classes]. Enschede, The Netherlands: University of Twente.

Eames, F. H. (1989). *A study of the effectiveness of instruction in multi-age grading vs. traditional single-grade level organization on the reading achievement of fourth graders.* Unpublished master's thesis, Western Connecticut State University. (ERIC Document Reproduction Service No. ED 309 388)

Fippinger, F. (1967). Empirische Untersuchung zur Leistung von Schülern aus voll und wenig gegliederten Schulen [An empirical study of scholastic achievement of pupils in fully and partially grade-differentiated schools]. *Schule und Psychologie, 14*(4), 97–104.

Fuller, B. (1987). What school factors raise achievement in the third world? *Review of Educational Research, 57,* 255–292.

Givens, H., Jr. (1972). *A comparative study of achievement and attitudinal characteristics of black and white intermediate pupils in individualized-multigrade and self-contained instructional programs.* Unpublished doctoral dissertation, Saint Louis University, Saint Louis, Missouri.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gutiérrez, R., & Slavin, R. E. (1992). Achievement effects of the nongraded elementary school: A best evidence synthesis. *Review of Educational Research, 62,* 333–376.

Harvey, S. B. (1974). *A comparison of kindergarten children in multigrade and traditional settings on self-concept, social-emotional development, readiness development, and achievement.* Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29–38). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hoen, R. R. (1972). *An evaluation of multi-age classes at Carnavon School, 1971–72.* Vancouver, British Columbia, Canada: Board of School Trustees, Department of Planning and Evaluation. (ERIC Document Reproduction Service No. ED 076 649)

Jarousse, J. P., & Mingat, A. (1991). *Les facteurs agissant sur les acquisitions des élèves à l'école primaire au Burkina Faso* [Factors influencing the achievement of primary school students in Burkina Faso]. Dijon, France: Université de Bourgogne, Institut de Recherche sur L'Economie de L'Education.

Jarousse, J. P., & Mingat, A. (1992). *L'Ecole primaire en Afrique: Fonctionnement, qualité, produits. Le cas du Togo* [The primary school in Africa: Function, quality, results. The case of Togo]. Dijon, France: Université de Bourgogne, Institut de Recherche sur L'Economie de L'Education.

Jokinen, H. (1979). *Yhdysluokkakoulujen oppilaiden koulusaavutuksista 2: Kouluhallituksen yhteisiin kokeisiin perustuva selvitys* [On the achievement of pupils in schools with combined grades: Part 2. A study based on centrally administered achievement tests arranged by the National Board of General Education]. Jyväskylä, Finland: University of Jyväskylä, Institute for Educational Research.

Junell, J. S. (1971). *An analysis of the effects of multigrading on a number of noncognitive variables.* Unpublished doctoral dissertation, University of Washington, Seattle.

Knight, E. E. (1938). A study of double grades in New Haven city schools. *Journal of Experimental Education, 7*(1), 11–18.

Knörzer, W. (1985). *Sind Schüler in kombinierten Grundschulklassen benachteiligt? Eine empirische Untersuchung* [Are students in multigraded schools put at a disadvantage? An empirical study]. Baltmannsweiler, Germany: Pädagogischer Verlag.

Knuver, J. W. M. (1993). *De relatie tussen klas- en schoolkenmerken en het affectief functioneren van leerlingen* [The relation between classroom and school characteristics and students' affective functioning]. Groningen, The Netherlands: RION, Instituut voor Onderwijsonderzoek.

Kral, M. (1995). *Effecten van schoolgrootte: Een onderzoek naar effectiviteitsverschillen tussen combinatieklassen en enkelvoudige klassen in het basisonderwijs* [Effects of school size: Effectiveness of combination classes and single-grade classes in primary education]. Nijmegen, The Netherlands: Instituut voor Toegepaste Sociale Wetenschappen.

Lem, P., Veenman, S. A. M., & Voeten, M. J. M. (1990). Zeitnutzung und Schulleistungen in Mehrstufenklassen. *Zeitschrift für Pädagogische Psychologie, 4*(1), 37–45.

Lundberg, I., & Linnakylä, P. (1993). *Teaching reading around the world.* Hamburg, Germany: International Association for the Evaluation of Educational Achievement.

MacDonald, P. A., & Wurster, S. R. (1974). *Multiple grade primary versus segregated first grade: Effects on reading achievement.* Tempe: Arizona State University. (ERIC Document Reproduction Service No. ED 094 336)

Martens, C. C. (1954). Educational achievements of eight-grade pupils in one-room rural and graded town schools. *Elementary School Journal, 54,* 523–525.

Mason, D. A., & Burns, R. B. (1995). Teachers' view of combination classes. *Journal of Educational Research, 89,* 36–45.

Mason, D. A., & Burns, R. B. (1996). "Simply no worse, and simply no better" may simply be wrong: A critique of Veenman's conclusion about multigrade classes. *Review of Educational Research, 66,* 307–322.

Miller, B. A. (1990). A review of the quantitative research on multigrade instruction. *Research in Rural Education, 7*(1), 1–8.

Miller, B. A. (1991). A review of the qualitative research on multigrade instruction. *Journal of Research in Rural Education, 7*(2), 3–12.

Nieminen, R. (1979). *Yhdysluokkakoulujen oppilaiden koulusaavutuksista 3: IEA-aineistoon perustuva selvitys kansakoulun oppilaiden menestymisestä luonnontiedon ja äidinkielen kokeissa* [On the achievement of pupils in schools with combined grades: Part 3. A study on primary school pupils' achievement in science and mother tongue based on material collected in the IEA study]. Jyväskylä, Finland: University of Jyväskylä, Institute for Educational Research.

Pawluk, S. T. (1992). *A comparison of the academic achievement of students in multigrade elementary classrooms and students in self-contained single-grade elementary classrooms.* Unpublished doctoral dissertation, Montana State University, Bozeman.

Pratt, D. (1986). On the merits of multiage classrooms. *Research in Rural Education, 3*(3), 111–115.

Rehwoldt, W., & Hamilton, W. (1957). *An analysis of some of the effects of interage and intergrade grouping in an elementary school.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Roelofs, E. (1993). *Teamgerichte nascholing en coaching: Een experimentele studie in scholen met combinatieklassen* [Staff development and coaching: An experimental study in schools with combination classes]. (Doctoral dissertation, University of Nijmegen). Nijmegen, The Netherlands: Universiteitsdrukkerij.

Roelofs, E., Raemaekers, J., & Veenman, S. (1991). Improving instructional and classroom management skills: Effects of a staff development programme and coaching. *School Effectiveness and School Improvement, 2*(3), 192–212.

Roelofs, E., Veenman, S., & Raemaekers, J. (1994). Improving instruction and classroom management behaviour in mixed-age classrooms: Results of two improvement studies. *Educational Studies, 20*(1), 105–126.

Rowley, S. D. (1992). *Multigrade classrooms in Pakistan: How teacher conditions and practices affect student achievement.* Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Schwarzer, R. (1991). Meta-Analysis Programs (Version 5.3) [Computer software]. Berlin, Germany: Freie Universität Berlin, Institut für Psychologie.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15*(9), 5–11.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57,* 293–336.

Spratt, B. R. (1986). *A comparative study of children enrolled in combination classes and non-combination classes in Fairfax County, Virginia Public Schools.* Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg.

Veenman, S. (1995). Cognitive and noncognitive effects of multigrade and multi-age classes: A best-evidence synthesis. *Review of Educational Research, 65,* 319–381.

Veenman, S., Lem, P., & Voeten, M. (1988). Time-on-task in mixed-age classes. *Journal of Classroom Interaction, 23*(2), 14–22.

Veenman, S., Lem, P., & Winkelmolen, B. (1985). Active learning time in mixed-age classes. *Educational Studies, 11*(3), 171–180.

Veenman, S., & Raemaekers, J. (1995). Long-term effects of a staff development programme on effective instruction and classroom management for teachers in multigrade classes. *Educational Studies, 21*(2), 167–185.

Veenman, S., Voeten, M., & Lem, P. (1987). Classroom time and achievement in mixed-age classes. *Educational Studies, 13*(1), 75–89.

Zabolotney, A. B. (1983). *A comparison of reading achievement and school attitudes of rural Seventh-Day Adventist multi-graded students and public school single-graded students in the state of Arkansas.* Unpublished doctoral dissertation, Brigham Young University, Provo, UT.

Author

SIMON VEENMAN is Senior Lecturer, Department of Educational Sciences, University of Nijmegen, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands; s.veenman@ped.kun.nl. He specializes in teacher education, staff development, and classroom instruction.

Received May 6, 1996
Accepted June 20, 1996